CLAIMS

What is claimed is:

1. A method of predicting the activity of a compound, the method comprising:

(a) providing a model for predicting activity of compounds based on one or more descriptors of physicochemical properties of compounds;

(b) characterizing the compound by one or more descriptor values;

(c) transforming the one or more descriptor values via one or more non-linear parametric functions, each of which provides an analytical relationship between the activity and the corresponding descriptor value;

(d) using the model to predict the activity of the compound from the transformed descriptors.

2. The method of claim 1, wherein the activity is a biological activity.

3. The method of claim 1, wherein the activity is a biological activity related to ADMET/Pk.

4. The method of claim 1, wherein the activity is a binding affinity to a biological molecule.

5. The method of claim 1, wherein the activity is binding to a cytochrome P450 enzyme.

6. The method of claim 1, wherein the compound is a therapeutic compound or a compound being investigated as potential therapeutic compound.

7. The method of claim 1, wherein at least one of the descriptors specifies a property of the compound as a whole.

8. The method of claim 7, wherein at least one of the descriptors is a lipophilicity-based descriptor of the compound, a size-based descriptor of the compound, or a charge-based descriptor of the compound.

9. The method of claim 1, wherein one of the non-linear parametric functions is a piece-wise continuous function representing a kernal-smoothed activity to descriptor relationship in training data.

10. The method of claim 1, wherein one of the non-linear parametric functions is a unimodal function.

11. The method of claim 1, wherein one of the non-linear parametric functions is a Gaussian function.

12. The method of claim 1, wherein one of the non-linear parametric functions is an asymptotic function.

13. The method of claim 1, wherein one of the non-linear parametric functions is a sigmoid function.

14. The method of claim 1, wherein one of the non-linear parametric functions is a hyperbolic function.

15. The method of claim 1, wherein the model provides a linear combination of a transformed descriptor value with one or more other transformed or non-transformed descriptor values.

16. The method of claim 1, wherein the model comprises a multiplicative combination of a first transformed descriptor value with one or more other transformed or non-transformed descriptor values.

17. The method of claim 16, wherein the model comprises a multivariate unimodal distribution.

18. The method of claim 16, wherein the model comprises a multivariate Gaussian function.

19. The method of claim 16, wherein the model the model comprises a multivariate asymptotic function.

20. The method of claim 16, wherein the model the model comprises a multivariate sigmoid function.

21. The method of claim 16, wherein the model the model comprises a multivariate hyperbolic function.

22. The method of claim 1, wherein (c)-(e) are performed in a single operation.

23. A computer program product comprising a machine readable medium on which is provided program instructions for predicting the activity of a compound, the program instructions comprising instructions for

(a) providing a model for predicting activity of compounds based on one or more descriptors of physicochemical properties of compounds;

(b) characterizing the compound by one or more descriptor values;

(c) transforming the one or more descriptor values via one or more non-linear parametric functions, each of which provides an analytical relationship between the activity and the corresponding descriptor value;

(d) using the model to predict the activity of the compound from the transformed descriptors.

24. The computer program product of claim 23, wherein the activity is a biological activity related to ADMET/PK.

25. The computer program product of claim 23, wherein the activity is a binding affinity to a biological molecule.

26. The computer program product of claim 23, wherein the activity is binding to a cytochrome P450 enzyme.

27. The computer program product of claim 23, wherein one of the non-linear parametric functions is a unimodal function.

28. The computer program product of claim 23, wherein one of the non-linear parametric functions is a Gaussian function.

29. The computer program product of claim 23, wherein one of the non-linear parametric functions is an asymptotic function.

30. The computer program product of claim 23, wherein the model comprises a multiplicative combination of a first transformed descriptor value with one or more other transformed or non-transformed descriptor values.

31. A method of predicting the activity of a compound, the method comprising:

(a) providing a model for predicting activity of compounds based on two or more descriptors of physicochemical properties of compounds, wherein the model comprises (i) a first non-linear transformation function which transforms values of a first descriptor and provides an analytical relationship between the activity and the first descriptor and (ii) a second non-linear transformation function which transforms values of a second descriptor and provides an analytical relationship between the activity and the second descriptor;

(b) characterizing the compound by a first descriptor value and a second descriptor value; and

(c) predicting the activity of the compound by using the first and second descriptor values as arguments in the model.

32. The method of claim 31, wherein at least one of the first and second descriptor values comprise a vector through descriptor space.

33. A computer program product comprising a machine readable medium on which is provided program instructions for predicting the activity of a compound, the program instructions comprising the following instructions:

(a) providing a model for predicting activity of compounds based on two or more descriptors of physicochemical properties of compounds, wherein the model comprises (i) a first non-linear transformation function which transforms values of a first descriptor and provides an analytical relationship between the activity and the first descriptor and (ii) a second non-linear transformation function which transforms values of a second descriptor and provides an analytical relationship between the activity and the second descriptor;

(b) characterizing the compound by a first descriptor value and a second descriptor value; and

(c) predicting the activity of the compound by using the first and second descriptor values as arguments in the model.

34. The computer program product of claim 33, wherein at least one of the first and second descriptor values comprise a vector through descriptor space.

35. A method of predicting the activity of a compound, the method comprising:

(a) providing a model for predicting activity of compounds based on two or more descriptors of physicochemical properties of compounds;

(b) characterizing the compound by the two or more descriptor values;

(c) generating one or more orthogonal, linear transformations of the descriptor values;

(d) transforming the linear transforms via one or more non-linear parametric functions, each of which provides an analytical relationship between the activity and the corresponding linear transform;

(e) using the model to predict the activity of the compound from the linearly transformed descriptors and subsequent non-linear transforms of the linearly transformed descriptors.

36. A method of creating a multivariate model for predicting the activity of compounds, the method comprising:

(a) for each compound for a plurality of compounds, obtaining an indication of the activity of the compound and values of one or more descriptors of the compound;

(b) for each of the descriptors, using activity versus descriptor data for the plurality of compounds to identify a non-linear parametric function representing an analytical relationship between the activity and the descriptor data;

(c) for each of the descriptors, transforming the descriptor via the non-linear parametric function; and

(d) creating the multivariate model of activity as a function of the descriptors and the non-linear parametric function transforms of the descriptors.

37. The method of claim 36, wherein the activity is a biological activity.

38. The method of claim 36, wherein the activity is a biological activity related to ADMET/PK.

39. The method of claim 36, wherein the activity is binding affinity to a biological molecule.

40. The method of claim 36, wherein the activity is binding to a cytochrome P450 enzyme.

41. The method of claim 36, wherein the plurality of compounds comprises at least one of (i) known therapeutic compounds and (ii) compounds being investigated as potential therapeutic compounds.

42. The method of claim 36, wherein at least one of the descriptors specifies a property of compounds as a whole.

43. The method of claim 42, wherein at least one of the descriptors is a lipophilicity-based descriptor, a size-based descriptor, or a charge-based descriptor.

44. The method of claim 36, wherein one of the non-linear parametric functions is a piece-wise continuous function representing a kernal-smoothed activity to descriptor relationship in a training data.

45. The method of claim 36, wherein one of the non-linear parametric functions is a unimodal function.

46. The method of claim 36, wherein one of the non-linear parametric functions is a Gaussian function.

47. The method of claim 36, wherein one of the non-linear parametric functions is an asymptotic function.

48. The method of claim 36, wherein one of the non-linear parametric functions is a sigmoid function.

49. The method of claim 36, wherein one of the non-linear parametric functions is a hyperbolic function.

50. The method of claim 36, wherein the model comprises a linear combination of a first non-linear transformed descriptor and one or more other transformed or non-transformed descriptors.

51. The method of claim 50, wherein the model comprises a linear combination of the first transformed descriptor and the one or more other transformed or non-transformed descriptor, each multiplied by a corresponding coefficient.

52. The method of claim 36, wherein the model comprises a multiplicative combination of the first transformed descriptor and the one or more other transformed or non-transformed descriptors.

53. The method of claim 52, wherein the model comprises a multivariate unimodal distribution.

54. The method of claim 52, wherein the model the model comprises a multivariate Gaussian function.

55. The method of claim 52, wherein the model the model comprises a multivariate asymptotic function.

56. The method of claim 52, wherein the model the model comprises a multivariate sigmoid function.

57. The method of claim 52, wherein the model the model comprises a multivariate hyperbolic function.

58. The method of claim 36, wherein creating the multivariate model comprises performing an optimization routine to determine the parameters for the non-linear parametric functions.

59. The method of claim 36, wherein (b)-(d) are performed together in a parameter optimization operation.

60. The method of claim 36, wherein the indication of the activity of the compound is a numerical value of activity.

61. The method of claim 36, wherein the indication of the activity of the compound is whether or not the compound possesses the activity.

62. The method of claim 36, further comprising, prior to (d), automatically identifying one or more descriptors as having little effect on the activity, and removing the one or more descriptors from use in the multivariate model.

63. A method of creating a model for predicting the activity of compounds, the method comprising:

(a) for each compound for a plurality of compounds, obtaining an indication of the activity of the compound and values of two or more descriptors of the compound; and

(b) creating the multivariate model of activity as a function of the first and second descriptors by using the values of the two or more descriptors and the indication of the activity for each of the plurality of compounds and fitting the descriptors and indications of activity to two or more non-linear parametric functions of the descriptors, which non-linear parametric functions are present in the model, wherein each of the non-linear parametric functions provides an analytical relationship between the activity and the corresponding descriptor value.

64. The method of claim 63, wherein creating the multivariate model comprises performing a minimization routine on descriptor and activity data from the plurality of compounds.

65. The method of claim 64, wherein the activity data comprises numerical values of activity.

66. The method of claim 64, wherein the activity data comprises an average expected numerical value of activity for at least some of the plurality of compounds.

67. The method of claim 66, wherein the activity data further comprises numerical values of activity for others of the plurality of compounds.

68. The method of claim 63, wherein the activity is binding to a CYP enzyme and at least one of the non-linear parametric transformation functions is a Gaussian function.

69. A method of creating a model for predicting the activity of a compound, the method comprising:

(a) for each compound for a plurality of compounds, obtaining an indication of the activity of the compound and values of two or more descriptors of the compound;

(b) generating one or more orthogonal, linear transformations of the descriptor values;

(c) for each of the linear descriptor transforms, using activity data versus the linear transform for the plurality of compounds to identify a non-linear parametric function, which provides an analytical relationship between the activity and the corresponding linear descriptor transform;

(d) for each of the linear descriptor transforms, transforming it via the non-linear parametric function; and

(e) creating the multivariate model of activity as a function of the linearly transformed descriptors and the subsequent non-linear transforms of these linearly transformed descriptors.

70. The method of claim 69, wherein generating the orthogonal linear transformations comprises applying a principal component analysis to the plurality of compounds.

71. A method of creating a multivariate model for predicting the activity of compounds, the model including at least one non-linear parametric transformation function that transforms a descriptor of the compounds, the method comprising:

(a) fitting activity versus descriptor data for a training set of compounds using an optimization function comprising (i) a penalty term that drives a parameter of the transformation function toward a boundary value, and (ii) an error term that compares a predicted activity with an actual activity for members of the training set;

(b) identifying one or more descriptors whose parameters were driven toward the boundary value by the penalty function to values below a threshold of significance to the model;

(c) eliminating the one or more descriptors identified in (b) from use with the multivariate model.

72. The method of claim 71, further comprising performing (a) – (c) at least twice.

73. The method of claim 71, further comprising fitting a model with the descriptors remaining after (c).

74. The method of claim 71, further comprising cross-validating a model using the descriptors remaining after (c).

75. The method of claim 74, further comprising de-selecting descriptors driven toward the boundary value during the cross-validation.

76. The method of claim 71, wherein the parameter of the transformation function driven by the penalty term is the width of the non-linear parametric transformation function for each descriptor.

77. The method of claim 76, wherein the penalty term includes terms representing the reciprocal of the width of the non-linear parametric transformation function.

78. The method of claim 71, wherein the non-linear parametric transformation function is a unimodal function.

79. The method of claim 71, wherein the non-linear parametric transformation function is a Gaussian function.

80. The method of claim 71, wherein the non-linear parametric transformation function is an asymptotic function.

81. The method of claim 71, wherein the non-linear parametric transformation function is a sigmoid function.

82. The method of claim 71, wherein the non-linear parametric transformation function is a hyperbolic function.

83. The method of claim 71, wherein (a)-(c) are performed automatically by a computing device, without user intervention.

84. The method of claim 74, wherein cross-validation is performed automatically by a computing device, without user intervention.

85. The method of claim 75, wherein de-selection of descriptors after cross-validation is performed automatically by a computing device, without user intervention.

86. The method of claims 74, wherein multiple rounds of model fitting, cross-validation, and descriptor de-selection are performed automatically by a computing device, without user intervention, until the process converges to a model based on a minimum set of descriptor identified as significant for the model.

87. The method of claim 71, wherein the optimization function further comprises scaling factors which specify the relative importance of the penalty terms with respect to the error term.

88. The method of claim 71, wherein the optimization function further comprises a parameter constraint term that promotes the numerical stability of the fit.

89. A computer program product comprising a machine readable medium on which is provided program instructions for creating a multivariate model for predicting the activity of compounds, the model including at least one non-linear parametric transformation function that transforms a descriptor of the compounds, the program instructions comprising instructions for

(a) fitting activity versus descriptor data for a training set of compounds using an optimization function comprising (i) a penalty term that drives a parameter of the transformation function toward a boundary value, and (ii) an error term that compares a predicted activity with an actual activity for members of the training set;

(b) identifying one or more descriptors whose parameters were driven toward the boundary value by the penalty function to values below a threshold of significance to the model;

(c) eliminating the one or more descriptors identified in (b) from use with the multivariate model.

90. The computer program product of claim 89, further comprising instructions for performing (a) – (c) at least twice.

91. The computer program product of claim 89, further comprising instructions for fitting a model with the descriptors remaining after (c).

92. The computer program product of claim 89, further comprising instructions for cross-validating a model using the descriptors remaining after (c).

93. The computer program product of claim 92, further comprising instructions for de-selecting descriptors driven toward the boundary value during the cross-validation.

94. The computer program product of claim 89, wherein the parameter of the transformation function driven by the penalty term is the width of the non-linear parametric transformation function for each descriptor.

95. The computer program product of claim 89, wherein the non-linear parametric transformation function is a unimodal function.

96. The computer program product of claim 89, wherein the non-linear parametric transformation function is a Gaussian function.

97. The computer program product of claim 89, wherein the non-linear parametric transformation function is an asymptotic function.

98. The computer program product of claim 89, wherein the non-linear parametric transformation function is a sigmoid function.

99. The computer program product of claim 89, wherein the non-linear parametric transformation function is a hyperbolic function.

100. The method of claim 1, wherein the activity is a binding affinity to a drug metabolizing enzyme.

101. The computer program product of claim 23, wherein the activity is a binding affinity to a drug metabolizing enzyme.

102. The computer program product of claim 23, wherein one of the non-linear parametric functions is a sigmoid function.

103. The computer program product of claim 23, wherein one of the non-linear parametric functions is a hyperbolic function.

104. The method of claim 36, wherein the activity is binding affinity to a drug metabolizing enzyme.